DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale

Year: 2022 | Citations: 414 | Authors: Samyam Rajbhandari, Conglong Li, Z. Yao, Minjia Zhang, Reza Yazdani Aminabadi

---

Abstract

As the training of giant dense models hits the boundary on the availability and capability of the hardware resources today, Mixture-of-Experts (MoE) models become one of the most promising model architectures due to their significant training cost reduction compared to a quality-equivalent dense model. Its training cost saving is demonstrated from encoder-decoder models (prior works) to a 5x saving for auto-aggressive language models (this work along with parallel explorations). However, due to the much larger model size and unique architecture, how to provide fast MoE model inference remains challenging and unsolved, limiting its practical usage. To tackle this, we present DeepSpeed-MoE, an end-to-end MoE training and inference solution as part of the DeepSpeed library, including novel MoE architecture designs and model compression techniques that reduce MoE model size by up to 3.7x, and a highly optimized inference system that provides 7.3x better latency and cost compared to existing MoE inference solutions. DeepSpeed-MoE offers an unprecedented scale and efficiency to serve massive MoE models with up to 4.5x faster and 9x cheaper inference compared to quality-equivalent dense models. We hope our innovations and systems help open a promising path to new directions in the large model landscape, a shift from dense to sparse MoE models, where training and deploying higher-quality models with fewer resources becomes more widely possible.